On the multibin logarithmic score used in the FluSight competitions

Bracher, Johannes

DOI: https://doi.org/10.1073/pnas.1912147116

Originally published at:
Bracher, Johannes (2019). On the multibin logarithmic score used in the FluSight competitions. Proceedings of the National Academy of Sciences of the United States of America, 116(42):20809-20810.
DOI: https://doi.org/10.1073/pnas.1912147116

On the multibin logarithmic score used in the FluSight competitions

Johannes Bracher

November 26, 2019

Epidemiology, Biostatistics and Prevention Institute, University of Zurich,
Hirschengraben 84, 8001 Zurich, Switzerland

johannes.bracher@uzh.ch

This is a preprint of a letter published in PNAS (https://doi.org/10.1073/pnas.1912147116). In their reply, Reich et al (https://doi.org/10.1073/pnas.1912694116) discuss the usefulness of different scoring rules in a public health context.

The *FluSight* challenges [9] represent an outstanding collaborative effort and have "pioneered infectious disease forecasting in a formal way" [10]. However, I would like to initiate a discussion about the employed evaluation measure.

The competitions feature discrete or discretized targets related to the US influenza season. *Eg* for the peak timing Y, a forecast distribution F consists of probabilities $p_1, \dots, p_T$ for the $T = 33$ weeks of the season. Such forecasts can be evaluated using the log score [2, 3]

$$\text{logS}(F, y_{\text{obs}}) = \log(p_{y_{\text{obs}}})$$

where y_{obs} is the observed value. This score is *strictly proper*, *ie* its expectation is uniquely maximized by the true distribution of Y. In the *FluSight* competitions the logS is applied in a *multibin* version,

$$\text{MBlogS}(F, y_{\text{obs}}) = \log\left(\sum_{i=-d}^{d} p_{y_{\text{obs}}+i}\right),$$

to measure accuracy of practical significance [9]. Depending on the target, d is either 1 or 5. Following the competitions, this score has become widely used [1, 5, 4, 6, 8, 7], even though as also mentioned in [9], it is improper. This may be problematic as improper scores incentivize dishonest forecasts. Assume $T > 2d$ and

$$p_1 = \dots = p_d = p_{T-d+1} = \dots = p_T = 0, \tag{1}$$

ie zero probabilities for the $2d$ extreme categories. Now define a *blurred* distribution $\tilde{F}$ with

$$\tilde{p}_t = \frac{\sum_{i=-d}^{d} p_{t+i}}{2d+1}, t = 1, \dots, T, \tag{2}$$

where $p_t = 0$ for $t < 1$ and $t > T$ and (1) ensures $\sum_{t=1}^{T} \tilde{p}_t = 1$. This implies

$$\text{MBlogS}(F, y_{\text{obs}}) = \text{logS}(\tilde{F}, y_{\text{obs}}) + \log(2d + 1),$$

ie the MBlogS is essentially the logS applied to a blurred version of F. To optimize the expected MBlogS under her true belief F, a forecaster should therefore not report F, but a sharper forecast G so that the blurred version $\tilde{G}$ (with $\tilde{p}_{G,1}, \dots, \tilde{p}_{G,T}$ derived from $p_{G,1}, \dots, p_{G,T}$ as in (2)) is close or equal to F. This follows from the propriety of the logS. An optimal G is found by maximizing $\sum_{t=1}^{T} p_t \cdot \log(\tilde{p}_{G,t})$ with respect to $p_{G,1}, \dots, p_{G,T}$.

This optimal G can differ considerably from the original F, as Fig. 1 shows for forecasts of the 2016/17 peak timing by the LANL team [8] (downloaded from `https://github.com/FluSightNetwork/cdc-flusight-ensemble/`). The optimized G (with $d = 1$) often have their mode shifted by one week and tend to be multimodal, even for unimodal F. Averaged over the 2016/17 season they yield improved MBlogS for the peak timing (-0.434 vs. -0.484). This illustrates that the MBlogS may be gamed, even though we strongly doubt participants have tried so. The logS, like any other proper score, could avoid such pitfalls.

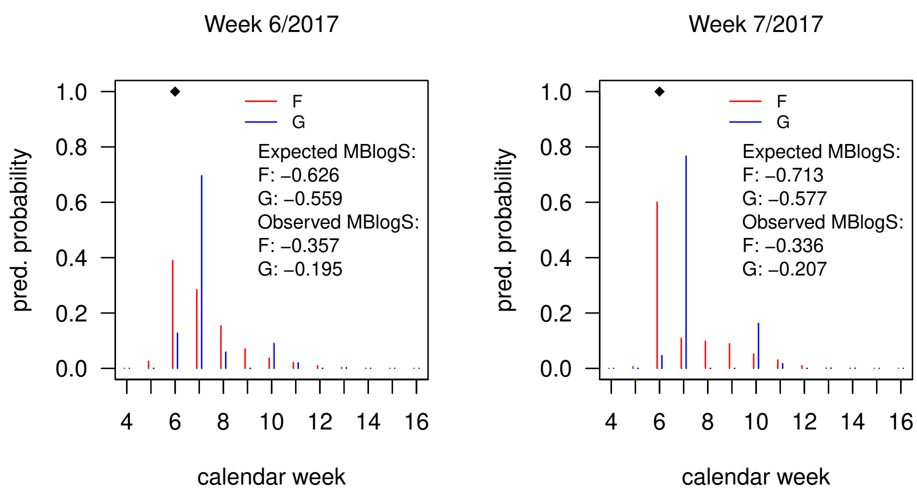

Figure 1: Forecasts F for the peak week, submitted by the LANL team in weeks 6–7, 2017, and optimized versions G. Diamonds mark the observed peak week. Expected scores are computed under F.

Acknowledgements: I would like to thank T. Gneiting for helpful discussions and the *FluSight Collaboration* for making its forecasts publicly available.

References

[1] Brooks, L. C., Farrow, D. C., Hyun, S., Tibshirani, R. J., and Rosenfeld, R. (2018). Nonmechanistic forecasts of seasonal influenza with iterative one-week-ahead distributions. *PLOS Comput Biol*, 14(6):1–29.

[2] Gneiting, T. and Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *J Am Stat Assoc*, 102(477):359–378.

[3] Held, L., Meyer, S., and Bracher, J. (2017). Probabilistic forecasting in infectious disease epidemiology: the 13th Armitage lecture. *Stat Med*, 36(22):3443–3460.

[4] Kandula, S. and Shaman, J. (2019). Near-term forecasts of influenza-like illness: An evaluation of autoregressive time series approaches. *Epidemics*, 27:41–51.

[5] Kandula, S., Yamana, T., Pei, S., Yang, W., Morita, H., and Shaman, J. (2018). Evaluation of mechanistic and statistical methods in forecasting influenza-like illness. *J Royal Soc Interface*, 15(144):20180174.

[6] McGowan, CJ et al (2019). Collaborative efforts to forecast seasonal influenza in the United States, 2015–2016. *Sci Rep*, Article Nr. 683.

[7] Osthus, D., Daughton, A. R., and Priedhorsky, R. (2019a). Even a good influenza forecasting model can benefit from internet-based nowcasts, but those benefits are limited. *PLOS Comput Biol*, 15(2):1–19.

[8] Osthus, D., Gattiker, J., Priedhorsky, R., and Del Valle, S. Y. (2019b). Dynamic Bayesian influenza forecasting in the United States with hierarchical discrepancy (with discussion). *Bayesian Anal*, 14(1):261–312.

[9] Reich, NG et al (2019). A collaborative multiyear, multimodel assessment of seasonal influenza forecasting in the United States. *Proc Natl Acad Sci*, 116(8):3146–3154.

[10] Viboud, C. and Vespignani, A. (2019). The future of influenza forecasts. *Proc Natl Acad Sci*, 116(8):2802–2804.